Message

From: Blancato, Jerry [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=232DE363DADB4CD9961900E10F56FDDF-BLANCATO, JERRY]
Sent: 1/24/2020 5:22:43 PM
To: Montilla, Alex [/o=ExchangeLabs/ou=Exchange Administrative Group (FYDIBOHF23SPDLT)/cn=Recipients/cn=b148b5335ff44aea8970035668052f01-Montilla, Alex]
Subject: Re: Requested Information

Thank you. Very good.

Jerry

Jerry N Blancato, PhD
Office of Research and Development
Senior Information Official

On Jan 24, 2020, at 11:09 AM, Montilla, Alex <Montilla.Alex@epa.gov> wrote:

Hi Jerry,

Please reference yesterday's IM chat request, *"...today at lunch we talked about a number of different initiatives we could do to manage data better, store data, and make data available along with governance."* Here is a summary of the topics we discussed.

- PFAS - The National PFAS Data Explorer obtains some of its data from WQP. This data is first loaded in WQX by the states and is uploaded using a spreadsheet template or an XML file. Once the data is received in WQX the chemical name is used to manually compare the new entry with SRS for lack of any unique identifier like the CASRN of the compound. This creates a heightened chance of human error in the reported PFAS data because of the numerous types of compounds and variations in their names, etc... There are approximately 7000 PFAS chemical in the Comp Tox TRI list. and 160 reportable PFAS chemicals in the most recent National Defense Authorization Act . We're attempting to learn about how SRS updates its index of PFAS chemicals. Once the submission is approved via the SRS verification process the WQP is uploaded with the data. At this point our ERG support contractor downloads the WQP file and manually reconciles it with the Comp Tox TRI list and the new NDAA list again introducing a chance for cross referencing errors. The idea is to create a public facing database for upload and entry of this water quality data applying frontend validation techniques a the point of data entry (e.g. Drop down lists of chemical names and/or list of CASRN) thus reducing the error rate and making reconciliation easier. I believe the offices of primary responsibility for this would be OW and OMS but this would help ORD tremendously.

- ScienceHub – The SH Dev Team and I met with DOE/OSTI to get a demonstration of their E-Link system. E-Link is a public facing portal that serves DOE intramural and extramural researchers. Manuscripts and data (to include software) are uploaded into E-Link and vetted for clearance and subsequent release to the public on the OSTI.gov website. I arranged the demonstration with the SH Dev Team because I have tasked them to begin researching the pros and cons of migrating EPA published datasets to EPA Data Commons on AWS as well as converting SH into a cloud application. E-Link serves as a

good example of how SH can evolve to serve EPA's open access efforts more broadly. The 3 short videos below provide a glimpse of what is in the realm of possibility if SH is expanded in the cloud.

- https://www.youtube.com/watch?v=xCzCNdxp3E0
- https://www.youtube.com/watch?v=Rbd80ANWLek
- https://www.youtube.com/watch?v=mZky964Lef8

- ORCID – We discussed becoming a member of the DOE/OSTI's ORCID Consortium which will save us money for current ORCID services and create opportunities to further expand our use of ORCID for data management purposes. For example, one stop shopping for profile information which can integrate with VIVO. Also, while we already make ORCID information available through SH for both ORD and non-ORD users we can envision integrating RAPID and/or STICS with ORCID information further enabling linkage of researcher information across systems.

- Datacite – We discussed enabling Pasteur and/or the EPA Data Commons to send consumption statistics to Datacite for their analysis and reporting. The goal here would be to better address the impact of our published data beyond number of downloads, hits and visits. Ultimately we want to know if our data has led to new discoveries, increased economic benefits and reduced lab to market times. With the reporting to Datacite the potential exists to leverage our metadata records to a greater degree for comparison and analysis with Crossref and ORCID information.

Lastly, all of this is in the queue to be addressed in FY2020. However, the looming RAPID/SH integration takes precedence before advancing any of these initiatives so it is best to get the RAPID/SH integration behind us sooner rather than later in order to move forward more quickly with these other initiatives. This is what I remember from yesterday's discussion. I hope this helps.

Thanks,

Alex